Detecting Argumentative Fallacies *in the Wild*: Problems and Limitations of Large Language Models

Ramon Ruiz-Dolz and **John Lawrence**
Centre for Argument Technology (ARG-tech)
University of Dundee
United Kingdom
{rruizdolz001,j.lawrence}@dundee.ac.uk

Abstract

Previous work on the automatic identification of fallacies in natural language text has typically approached the problem in constrained experimental setups that make it difficult to understand the applicability and usefulness of the proposals in the real world. In this paper, we present the first analysis of the limitations that these data-driven approaches could show in real situations. For that purpose, we first create a validation corpus consisting of natural language argumentation schemes. Second, we provide new empirical results to the emerging task of identifying fallacies in natural language text. Third, we analyse the errors observed outside of the testing data domains considering the new validation corpus. Finally, we point out some important limitations observed in our analysis that should be taken into account in future research in this topic. Specifically, if we want to deploy these systems *in the Wild*.

1 Introduction

In the field of the automatic analysis of natural language argumentative discourse, the identification of fallacies plays an important role since it can be a determining feature to measure the *quality* of argumentation (Wachsmuth et al., 2017). Furthermore, the automatic identification of fallacies can also be helpful for the development of disinformation detection systems and critical thinking tools (Visser et al., 2020). Studied since the times of the ancient Greece by Aristotle (Aristotle, 1978), a fallacy was seen as an argumentation strategy used to deceive an opponent in a debate and unfairly get the reason. This definition evolved with time (Van Eemeren and Grootendorst, 1984; Hamblin, 1970) extending the instrumental notion of the Aristotelian fallacy to more modern theories of logic and mathematics. A more recent (and complete) definition was provided by Walton (1995), where fallacies are defined as "*important, baptizable types of errors or deceptive tactics of argumentation that tend to fool or trip up participants in argumentation in various kinds of everyday discussions*". This definition is less constrained and more accurate to the natural language challenges we may face these days.

Detecting a piece of fallacious reasoning, however, is not trivial and requires knowledge in a broad number of areas that make this task challenging. First, it is important to be able to analyse the logical reasoning underlying natural language arguments. For that purpose, it is required to distil the abstract and formal components from the informal natural language argument. This first case is that of *formal* fallacies (Oliver, 1967). Second, solid knowledge on the domain of discussion is of utmost importance. An argument can be logically sound but still fallacious, such is the case of *informal* fallacies (Walton, 1987). Therefore, only with a complete analysis is it possible to determine if a natural language argument is a fallacy or not, as well as the underlying reasons why it is fallacious. A consistent way to conduct this analysis is to rely on validated models of argument which capture the notion of fallacy. Different models have been proposed and studied in the literature; such is the case of the pragma-dialectic theory of argumentation (Van Eemeren and Grootendorst, 2016) in which the authors define ten rules to guide argumentative discussions. The fulfilment of these rules allows to create a fruitful discussion, but an argument that breaks any of these rules is considered a fallacy. Another good example of these models is the argumentation schemes proposed by Walton (Walton et al., 2008). An argumentation scheme combines the abstract representation of the underlying logic of a natural language argument with a set of critical questions that must be successfully answered to prove the validity of an argument. The argumentation scheme model is very interesting w.r.t. fallacy analysis, since an argument being fallacious is not determined by belonging to a specific class, but depending on the answers provided to the set of

Proceedings of the 10th Workshop on Argument Mining, pages 1–10
December 7, 2023. ©2023 Association for Computational Linguistics

critical questions. For example, a natural language argument belonging to the *Ad Hominem* scheme is not a fallacy per se, but it must be structured as follows:

Character Attack Premise: a is a person of bad character.

Conclusion: a's argument α should not be accepted.

And it is only considered to be fallacious if any of the following critical questions cannot be successfully answered,

- CQ1: How well supported by evidence is the allegation made in the character attack premise?
- CQ2: Is the issue of the character relevant in the type of dialogue in which the argument was used?
- CQ3: Is the conclusion of the argument that α should be rejected, or is the conclusion that α should be assigned a reduced weight of credibility?

Therefore, with the argumentation scheme paradigm, it is possible to partially dissociate the natural language and the logic of the argument, allowing for a more informed analysis of the reasons of an argument being fallacious.

In this paper, we integrate the concept of argumentation schemes in the evaluation of machine learning and Transformer-based language models for the automatic detection of fallacies in natural language arguments. It is our objective to understand the way these models, as they have been proposed in most of the previous work in this topic, are able to *learn* the reasons behind a fallacy and generalise to data outside of the training domain. Our contribution is therefore threefold: (i) we create a fallacy validation corpus consisting of natural language argumentation schemes; (ii) we provide new empirical results for the emerging task of identifying fallacies in natural language text; and (iii) we analyse the observed errors inside and outside of the testing data domains considering the argumentation scheme validation corpus, and point out some of the main limitations of relying exclusively on LLMs when addressing complex natural language reasoning problems.

2 Related Work

The automatic detection of fallacies in natural language texts is an emerging topic of research within the area of Natural Language Processing. One of the first efforts in developing a database of fallacies was done in (Habernal et al., 2017) creating "*Argotario*", an educative platform where participants could improve their debating skills. Through gamification, the authors collected fallacies registered by the participants belonging to one of the following five classes: *ad hominem*, appeal to emotion, red herring, hasty generalisation, irrelevant authority. A direct continuation of this work was presented in (Habernal et al., 2018a), where the resulting corpus from the use of "*Argotario*" containing 430 annotated arguments was released. In that work, arguments belonging to the previous five classes plus a *no fallacy* set of arguments were compiled, and a set of preliminary results of experiments with a Support Vector Machine (SVM) and a Bidirectional Long Short-Term Memory (BiLSTM) neural network were reported.

Aimed at better understanding the linguistic features underlying the *Ad Hominem* argument, Habernal et al. developed a corpus from user discussions in the *Change My View* subreddit on the Reddit social network (Habernal et al., 2018b). For that purpose they retrieved the comments that were removed by the administrators because they were labelled as rude or hostile by the community, matching one of the non breakable rules proposed in (Van Eemeren and Grootendorst, 2016) as part of the pragma-dialectic theory of argumentation. The authors also reported a set of fallacy detection experiments with a Convolutional Neural Network (CNN) in which they used this corpus consisting of 7,242 samples balanced between non-fallacious and *ad hominem* classes.

The automatic identification of argumentative fallacies has also been studied from the propaganda viewpoint in (Da San Martino et al., 2019), where the authors annotate news articles containing up to 18 propaganda techniques and report a series of experiments on propaganda classification. This perspective on fallacious argumentation was continued in a shared task organised for the SemEval forum (Da San Martino et al., 2020) aimed at the automatic classification of natural language propaganda.

Based on the pragma-dialectic theory (Van Eemeren and Grootendorst, 2016) eight classes of fallacious arguments were annotated in a corpus of informal fallacies in online discussions by Sahai et al. (2021). More than 1,700 fallacious comments retrieved from Reddit were annotated into the classes of Appeal to authority, Appeal to majority, Appeal to nature, Appeal to tradition, Appeal to worse problems, Black-or-white, Hasty generalisation, and Slippery slope fallacies. Furthermore, the authors report results on the binary task of classifying natural language text as fallacious or not, and on the 8-class classification problem of determining the type of fallacy to which each fallacious comment belongs to. For the experiments, the authors consider more advanced models based in the Transformer architecture, and the granularity network that performed the best in (Da San Martino et al., 2019).

A simplified version of the task is presented in (Goffredo et al., 2022), where another corpus of fallacious argumentation is released. In this paper, the annotation of fallacious arguments is done from the transcripts of 31 political debates of the U.S. Presidential Campaigns. The authors annotate six different types of fallacy: *Ad Hominem*, Appeal to Emotion, Appeal to Authority, Slippery Slope, False Cause, and Slogans. In addition to these classes, 11 sub-classes are also annotated, providing additional information of the fallacious arguments. In their experiments, the best results are reported with a Transformer-based architecture that combines natural language with argumentative features. The experimental results reported in that work are exclusively focused on the task of classifying fallacies, assuming that the fallacy has already been detected.

Recently, (Alhindi et al., 2022) al explores the use of multitask instruction-based prompting to dectect 28 different fallacies across five datasets. The authors compare the use of T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020) for prompt-based fallacy classification, and a fine-tuned BERT (Devlin et al., 2018) model for a more classic baseline. From their results, it is possible to observe how the multitask instruction-based prompting with T5 achieves a significant increase in performance compared to the GPT-3 and BERT baselines. However, the methodology applied in this paper is similar to the one followed in previous work, in which the fallacy type of a text sequence is determined by only taking the natural language of the sequence into account.

Finally, one of the most recent papers in the automatic detection of logical fallacies proposes a new task for pre-training language models based on the structure of arguments (Jin et al., 2022). For that purpose, the authors release a corpus consisting of 2,449 argumentative samples labelled into one of 13 different fallacy types. A set of experiments comparing Large Language Models (LLMs) as zero-shot classifiers with Transformer-based models fine-tuned on the corpus is reported, emphasising on the importance of looking at structural reasoning features for this type of classification problems.

We can observe how, in the past years, a varied set of relatively small corpora have been annotated and publicly released. Most of them, however, share a similar paradigm for addressing the automatic identification of argumentative fallacies. Short spans of text are labelled with one of the corresponding fallacy labels, but no attention is given to the underlying logic that makes the argument fallacious or not. Furthermore, all the reported experiments are done in a similar way, the natural language text is used as the input to learn a set of N classes (varying from one corpus to another) directly from the text, and no in-depth error analyses are reported in most of these works. These limitations might raise some concerns, such as the impact of non-fallacious arguments being labelled as fallacious (false positives) while they are not, just because they share similar words or natural language patterns. To have a better understanding of these cases, and the potential problems of relying only in deep learning algorithms for addressing a complex problem such as the identification of natural language fallacies, the argumentation scheme model of arguments presents itself as a promising alternative to the models considered in the literature.

3 Data

In order to validate our hypothesis and to provide an evaluation outside of the training domain, we decided to use two different corpora in our experiments. First, the fallacy detection corpus, which consists of a partial combination of the data described in (Sahai et al., 2021) and (Goffredo et al., 2022). Second, the argumentation scheme validation dataset, a small collection of natural language argumentation schemes that we created in this work in order to evaluate the inferences done by the predictive models to detect a natural language fallacy outside of the domains considered during training. With this second dataset, it is our objective to observe how well the model generalises when detecting natural language fallacies following a different model or structure than the one considered in the data used for training, similar to what would happen when deploying the predictive models *in the Wild*.

As depicted in Table 1, the fallacy detection corpus used in this work consists of four fallacy classes and the non-fallacious class. We selected the fallacy classes of Appeal to Authority, Appeal to Majority, Slippery Slope and *Ad Hominem* since they represent the majority of the natural language fallacies commonly used in human dialogues and debates.

Since the annotation in both corpora was based on similar fallacy theory, our fallacy detection corpus combines some of the natural languages fallacies annotated in U.S. presidential debates (Goffredo et al., 2022), with some others annotated in social media discussions (Sahai et al., 2021) and the non-fallacious class. The decision of combining both corpora is twofold. First, we wanted to address the automatic detection of natural language fallacies (not just classifying them as done in (Goffredo et al., 2022)) so non-fallacious samples were needed. The assumption done in (Goffredo et al., 2022) of knowing beforehand that some piece of natural language is fallacious represents a significant limitation of the contribution since knowing the fallacious condition of an input is not trivial, and represents an important challenge in the area. The second reason to combine both corpora is to have a more balanced distribution of samples when comparing fallacious to non-fallacious samples, and to expand the natural language domains in which fallacies can be observed during training.

A sample in our fallacy detection corpus consists of a short snippet of text where the fallacious (or not) reasoning has been identified, a natural language context in which the fallacy has been detected (a paragraph in the case of the debates, and the previous comment of the text snippet in the case of the social media discussions), and the annotated label. In order to homogenise the natural language context in data belonging to both corpora, for the samples extracted from the debate corpus we considered as the context only the sentences before and after the text snippet.

Aimed at validating the performance of machine learning and deep learning systems to detect natural language fallacies, we developed a small dataset containing natural language argumentation schemes (Walton et al., 2008). In this dataset, we included seven different types of argumentation schemes matching the fallacy classes included in the fallacy detection corpus: Argument from Expert Opinion (AFEO), Argument from Position to Know (AFPK), Argument from Popular Practice (AFPP), Argument from Popular Opinion (AFPO), Slippery Slope Argumentation Scheme (SSAS), Generic *Ad Hominem* (GAH), and Circumstantial *Ad Hominem* (CAH). This way, we can easily relate each argumentation scheme with one of the four fallacy classes included in the fallacy detection corpus, the Appeal to Authority with AFEO and AFPO, the Appeal to Majority with AFPP and AFPO, the Slippery Slope with the SSAS, and the *Ad Hominem* with the GAH and CAH. It is important to remember that, argumentation schemes are not fallacious by definition as they are the fallacy classes used to annotate previous corpora, but they can only be considered as fallacious if and only if some of the critical questions cannot be successfully answered. Taking this into consideration, in our argumentation scheme validation dataset, we included two natural language instances of each scheme, one in which all the critical questions can be answered (i.e., valid reasoning), and another that fails in some aspect (i.e., fallacy). Therefore, our argumentation scheme validation dataset consists of fourteen natural language arguments specifically designed to validate the inference process of the predictive models in the task of automatically detecting natural language fallacies. These natural language argumentation schemes have been compiled in Table 2.

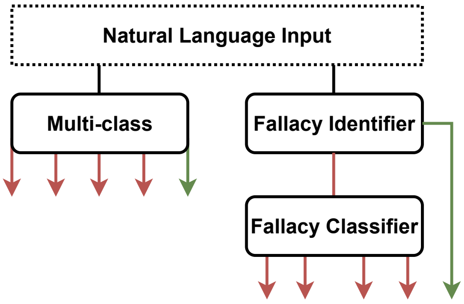

Figure 1: Multi-class and cascaded approaches.

Samples	Appeal to Authority	Appeal to Majority	Slippery Slope	Ad Hominem	Fallacy Total	Non-fallacious
(Sahai et al., 2021)	212	196	228	-	636	1650
(Goffredo et al., 2022)	208	-	48	146	402	-
Total	420	196	276	146	1038	1650

Table 1: Class distribution of the fallacy detection corpus.

4 Experiments

4.1 Method

To extend the experimental results previously reported in the literature, we consider the two different approaches to the automatic detection of argumentative fallacies depicted in Figure 1. First, we consider a multi-class classification problem in which fallacy classes and the non-fallacious class are considered in the same level. In this case, we will be facing a five-class classification problem. Second, we consider a cascaded approach in which we first try to discriminate fallacies from valid reasoning. For that purpose, we combine a two-class classifier in charge of detecting fallacies, with a four-class classification model that determines the specific type of the fallacy (i.e., Authority, Majority, Slipepry Slope, and *Ad Hominem*).

4.2 Experimental Setup

In our experiments, we have considered three different implementations of the fallacy classifiers proposed in our method. Aimed at covering some of the state-of-the-art general approaches in NLP, we used a Support Vector Machine combined with natural language embeddings (eSVM), a fine-tuned RoBERTa for sequence classification, and zero-shot prompting GPT-3.5-TURBO and GPT-4 without any additional training. We also considered two versions of each input in our experiments: (i) we used as our input the text snippet only, and (ii) we combined the snippet with its context.

Regarding the eSVM, the best results were obtained with the radial basis function kernel, a *gamma* equal to one divided by the number of features, and C equal to 1000. On the other hand, for fine-tuning the RoBERTa model, we trained the model for 20 epochs with a learning rate of 1e-5 and a weight decay of 0.01. Finally, the prompt used in our experiments with GPT-3.5-TURBO and GPT-4 to automatically detect and classify natural language fallacies was designed in three sequential messages as follows:

- *You task is to detect a fallacy in the Text Snippet. The label can be "Slippery Slope", "Appeal to Authority", "Ad Hominem", "Appeal to Majority" or "None".*
- *Text Snippet:* [SAMPLE]
- *Label:*

The first paragraph of the prompt was adapted for each of the different situations proposed in our method. For example by removing *"None"* for fallacy classification (4-class), and grouping the fallacy labels into *"Fallacy"* for fallacy identification (2-class).

In all of our experiments, we considered an 80-10-10 split of our data into train, development, and test respectively. Furthermore, we removed all the duplicated text snippets from the U.S. presidential debates corpus to prevent the occurrence of the same natural language snippets in train and test at the same time, as happened in the experiments reported in (Goffredo et al., 2022). The best performing hyperparameters described above were selected based on the best performance in the development split. The code and the data used in our experiments can be publicly accessed at https://github.com/raruidol/ArgumentMining23-Fallacy.

5 Results

We have grouped the analysis of our results into two sections. First, we evaluate our models on the test split of the fallacy detection corpus. Second, we evaluate these same models when used to detect or classify fallacies *in the Wild* (i.e., outside of the training/testing data domain), for which purpose we use the argumentation scheme validation dataset.

5.1 Experimental Evaluation

Regarding the experimental evaluation, we measured the performance of the models by calculating the precision, recall, and macro f1 of the predictions done over the test samples. Table 3 contains the results of the multi-class classification experiments, Table 4 contains the results of the fallacy

Arg. Scheme	CQs	Natural Language Argumentation Schemes
AFEO	✓	*Major Premise*: “Prof Whittaker is a professor of virology at the Cornell University College” *Minor Premise*: “Prof Whittaker said that viruses can be spread by sneezing” *Conclusion*: “Viruses can be spread by sneezing”
AFEO	✗	*Major Premise*: “Stephen Hawking was an expert on AI” *Minor Premise*: “Stephen Hawking said that AI could spell the end of the human race” *Conclusion*: “AI could spell the end of the human race”
AFPK	✓	*Major Premise*: “Alice lives in New York” *Minor Premise*: “Alice says that New York City Hall is in Lower Manhattan” *Conclusion*: “New York City Hall is in Lower Manhattan”
AFPK	✗	*Major Premise*: “David is a cab driver in London” *Minor Premise*: “David says that the best way to get to Tower Bridge is by cab” *Conclusion*: “The best way to get to Tower Bridge is by cab”
AFPP	✓	*Major Premise*: “Most people wear black clothes at a funeral” *Minor Premise*: “If most people wear black clothes at a funeral, that is acceptable to do” *Conclusion*: “It is acceptable to wear black clothes at a funeral”
AFPP	✗	*Major Premise*: “Most people drive at least 10 miles per hour over the speed limit” *Minor Premise*: “If most people drive at least 10 miles per hour over the speed... ...limit, that is acceptable to do” *Conclusion*: “It is acceptable to drive at least 10 miles per hour over the speed limit”
AFPO	✓	*General Acceptance Premise*: “The majority of climate scientists agree that humans... ...are causing global warming and climate change” *Presumption Premise*: “If the majority of climate scientists agree that humans... ...are causing global warming and climate change, there is a reason to believe that is true” *Conclusion*: “There is reason to believe that humans... ...are causing global warming and climate change”
AFPO	✗	*General Acceptance Premise*: “The majority of people we asked agreed that the Earth may be flat ” *Presumption Premise*: “If the majority of people we asked agreed that the Earth... ...may be flat, there is a reason to believe that is true” *Conclusion*: “There is reason to believe that the Earth may be flat”
SSAS	✓	*First Step Premise*: “I should go out with my friends rather than study for the exam” *Recursive Premise*: “If I don’t pass the exam, this might affect my GPA, which... ...in turn might impact my chances of going to a good college” *Bad Outcome Premise*: “Not going to a good college would be a disaster” *Conclusion*: “I should not go out with my friends rather than study for the exam”
SSAS	✗	*First Step Premise*: “We should lower the legal drinking age from 21 to 18 in line with other countries” *Recursive Premise*: “If we lower it to 18, next it will be 17, then 16, 15, etc. ” *Bad Outcome Premise*: “If we lower the legal drinking age, we’ll have ten-year-olds getting drunk in bars!” *Conclusion*: “We should not lower the legal drinking age ”
GAH	✓	*Character Attack Premise*: “Steve has cheated on a number of past exams” *Conclusion*: “We should doubt Steve’s claim that someone else copied his work in this exam”
GAH	✗	*Character Attack Premise*: “The CEO was convicted of a DUI in college” *Conclusion*: “We should doubt the CEO’s sales report”
CAH	✓	*Argument Premise*: “The car salesman argued that I should buy a gas car because... ...they are more reliable than electric cars” *Inconsistent Commitment Premise*: “The car salesman chose to drive an electric car” *Credibility Questioning Premise*: “The car salesman is not credible in this case” *Conclusion*: “The car salesman’s argument that I should buy a gas car is not valid”
CAH	✗	*Argument Premise*: “Mark argued that you should not take illegal drugs as they can have dangerous side effects” *Inconsistent Commitment Premise*: “Mark has taken illegal drugs in the past” *Credibility Questioning Premise*: “Mark is not credible in this case” *Conclusion*: “Mark’s argument that you should not take illegal drugs is not valid”

Table 2: Argumentation Scheme validation dataset. A (✓) indicates that the argument successfully answers its critical questions. A (✗) indicates that some of the critical questions cannot be successfully answered and thus, the argument is a fallacy.

Model	**Precision**	**Recall**	**Macro-F1**
RB	21.6	24.6	18.6
eSVM	68.3	55.8	60.3
RoBERTa	**68.2**	**65.3**	**66.5**
GPT-3.5-TURBO	59.0	46.2	45.5
GPT-4	53.5	55.0	51.7
eSVM+[*ctx*]	67.3	50.0	54.4
RoBERTa+[*ctx*]	62.0	58.4	59.9
GPT-3.5-TURBO+[*ctx*]	50.2	32.1	35.8
GPT-4+[*ctx*]	54.4	51.2	50.8

Table 3: Precision, Recall and Macro-F1 results of the 5-class fallacy detection task. [*ctx*] represents the contextual information added to the input of each model.

detection (i.e., 2-class classification) experiments, and Table 5 contains the results of the fallacy (i.e., 4-class) classification experiments. We have also included the random baseline (RB) in order to relativise the results with respect to the class complexity of each instance of the task. From all these results, we have identified two interesting patterns.

First of all, for a corpus of this size (i.e., ∼2000 samples) and distribution, the best results were consistently achieved by fine-tuning the RoBERTa architecture. The eSVM model performed slightly worse and the worst performing approach was the zero-shot prompts for the GPT-3.5-TURBO and GPT-4 model. It is important to mention that in the zero-shot prompting experiments, no parameters were specifically fine-tuned for our data, and taking this into account, the results were surprisingly good compared to a random or a majority baseline. Furthermore, we could observe an important difference between GPT-3.5-TURBO and GPT-4 when prompted to detect and classify fallacies in natural language. We found out that in all of the fallacy detection and classification tasks GPT-4 significantly outperformed GPT-3.5-TURBO. Specifically in the cascaded approach, GPT-4 was able to outperform GPT-3.5-TURBO in more than a 20% with respect to macro F1 reaching a maximum improvement of a 58% in the fallacy classification task. After removing the negative samples, the GPT-4 model is able to focus on more relevant linguistic aspects of the text snippets than its predecessor, resulting in a significant improvement in this task (see Table 5). Finally, we were also able to observe that in general, better results were achieved by the cascaded approach. Therefore, when addressing a fallacy identification problem, given the linguistic complexity of this task, it is better to do it by separating the detection and the classification than doing both tasks at the same time.

The second pattern that we were able to observe is that, including the context as we did in our experiments was not helpful at all. Adding more contextual information to the text snippet resulted in redundant information that made the task more difficult for the predictive models. Given the generalised bad performance of the models when just including the adjacent text of the snippet to the input, we consider that argumentative context should be brought into consideration from a different perspective (e.g., explicitly modelling the underlying reasoning of the argument). Since the detection of fallacious reasoning is a task that involves the analysis of finer grained reasoning and logical aspects of natural language, it might be a better idea to support the natural language input with some structural and argumentative features in the line of what was proposed in (Jin et al., 2022), rather than just including the adjacent text. However, we could not integrate such features in our experiments since part of the fallacy detection corpus did not contain such annotations. Finally, we would also like to point out that from the consistent drop of performance observed between all of our experiments with and without context, the development of an effective segmentation algorithm that focuses on the relevant linguistic aspects of the text is of utmost importance when addressing a high linguistic complexity task such as the automatic detection of argumentative fallacies.

5.2 Evaluation *in the Wild*

In order to validate the behaviour of these models when making predictions outside of the training domains, we have used the validation dataset created on the basis of the argumentation scheme model of argument (see Table 2). For this validation *in the Wild*, we have selected the best model of the experimental evaluation considering both fine-tuning and prompt-based models independently. As depicted in Table 6, we have evaluated the RoBERTa and GPT-4 models considering both the multi-class and the fallacy identification tasks (i.e., 5-class and 2-class classification problems respectively) proposed at the beginning of this paper.

Firstly, looking at the 5-class classification re-

Model	Precision	Recall	Macro-F1
RB	47.1	47.0	46.4
eSVM	77.8	77.5	77.7
RoBERTa	**79.8**	**79.6**	**79.6**
GPT-3.5-TURBO	41.7	46.2	40.6
GPT-4	53.2	53.2	51.1
eSVM+[*ctx*]	76.8	74.0	74.8
RoBERTa+[*ctx*]	78.0	78.8	78.3
GPT-3.5-TURBO+[*ctx*]	47.1	48.8	43.5
GPT-4+[*ctx*]	56.6	56.7	54.1

Table 4: Precision, Recall and Macro-F1 results of the 2-class fallacy detection task. [*ctx*] represents the contextual information added to the input of each model.

Model	Precision	Recall	Macro-F1
RB	22.9	22.1	22.4
eSVM	69.6	65.5	67.1
RoBERTa	75.4	**78.0**	**76.2**
GPT-3.5-TURBO	51.7	46.4	44.6
GPT-4	60.4	60.0	58.3
eSVM+[*ctx*]	**79.7**	72.1	74.8
RoBERTa+[*ctx*]	72.3	72.6	72.3
GPT-3.5-TURBO+[*ctx*]	45.9	38.1	35.1
GPT-4+[*ctx*]	58.7	57.0	55.7

Table 5: Precision, Recall and Macro-F1 results of the 4-class fallacy classification task. [*ctx*] represents the contextual information added to the input of each model.

Arg. Scheme	**CQs**	**RoBERTa**		**GPT-4**	
		5-class	**2-class**	**5-class**	**2-class**
AFEO	✓	Authority	Fallacy	None	None
AFEO	✗	Authority	Fallacy	Authority	Fallacy
AFPK	✓	None	Fallacy	None	None
AFPK	✗	Authority	Fallacy	Authority	Fallacy
AFPP	✓	None	Fallacy	Majority	None
AFPP	✗	None	Fallacy	Majority	Fallacy
AFPO	✓	Majority	Fallacy	Authority	None
AFPO	✗	Majority	Fallacy	Majority	Fallacy
SSAS	✓	None	Fallacy	Slippery Slope	None
SSAS	✗	Slippery Slope	Fallacy	Slippery Slope	Fallacy
GAH	✓	None	None	Ad Hominem	Fallacy
GAH	✗	Ad Hominem	Fallacy	Ad Hominem	Fallacy
CAH	✓	None	None	Ad Hominem	Fallacy
CAH	✗	None	None	Ad Hominem	Fallacy

Table 6: Evaluation *in the Wild* of the fallacy detection LLMs.

sults, we can observe different behaviour between RoBERTa and GPT-4. In the case of RoBERTa, it failed to distinguish the fallacious aspects of the underlying logic of four argumentation schemes. We can see this problem with both AFEO that are classified as an authority fallacy, both AFPP that are classified as non-fallacious while both AFPO are labelled as an appeal to majority fallacy, and both CAH that are classified as non-fallacious. This behaviour can be attributed to the fact that they look too similar to the samples labelled as fallacious (in the case of AFEO and AFPO) or non-fallacious (in the case of AFPP and CAH) in the training corpora. Only for three out of the seven argumentation schemes was the model able to correctly distinguish between fallacious and non-fallacious instances of the same scheme, this is the case of AFPK, SSAS, and GAH. Differently, GPT-4 only managed to correctly distinguish between an instance of the same argumentation scheme being fallacious or not in the AFEO and AFPO. All the rest of the argumentation schemes were labelled as fallacious belonging to each of its respective fallacy classes. It is interesting to mention that GPT-4 also failed to identify the fallacy type in the valid AFPO, since the word "*scientist*" appeared, the model predicted that it was an appeal to authority fallacy, being it not a fallacy and being structured as a popular opinion scheme, meaning that the authority was not a relevant aspect in the argumentative reasoning.

Secondly, looking at the 2-class classification results, the observed behaviour between RoBERTa and GPT-4 was also significantly different. In the case of RoBERTa, except for the *Ad Hominem* schemes, all the other argumentation schemes were labelled as fallacious regardless of their logic. The model was also not able to correctly discriminate a fallacy in the case of CAH arguments, where both of them were labelled as non-fallacious. Only the natural language GAH schemes were correctly discriminated between fallacious or not. On the other hand, GPT-4 performed surprisingly well in this instance of the task. All the schemes apart from the *Ad Hominem* ones were correctly classified as fallacious or not. However, both GAH and CAH schemes were labelled as fallacious, regardless of the actual reasons (e.g., critical questions) of being fallacious.

6 Discussion

In this paper, we present the first analysis of the limitations of approaching the fallacy detection problem with LLMs. For that purpose, we provide a new viewpoint to the existing work done in the automatic identification of natural language fallacies through the use of the argumentation scheme model of arguments. The argumentation scheme model allows us to partially dissociate the logic of the argument from the natural language of it, evidencing the limitation that LLMs have when used to approach complex natural language tasks where logical reasoning is involved. For that purpose, we first ran a set of experiments training a machine learning and a deep learning algorithm plus prompting two LLMs on existing annotated corpora for fallacy identification, resulting in new baselines for this task. Second, we evaluated the best performing models on a specifically created argumentation scheme validation dataset that helped us to understand how well were these models able to identify fallacies based on the logic of the argument rather than over-fitting to a natural language pattern not relevant for the definition of a fallacy. From our findings we have been able to observe that there is still much more work to do in this area, and that relying exclusively on LLMs to approach such a challenging task *in the Wild* may not be the best option.

Acknowledgements

This work has been supported by the 'AI for Citizen Intelligence Coaching against Disinformation (TITAN)' project, funded by the EU Horizon 2020 research and innovation programme under grant agreement 101070658, and by UK Research and innovation under the UK governments Horizon funding guarantee grant numbers 10040483 and 10055990.

References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aristotle. 1978. *De Sophisticis Elenchis (On Sophistical Refutations)*. Harvard University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

G Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}, International Joint Conferences on Artificial Intelligence Organization*, pages 4143–4149.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396.

Charles L Hamblin. 1970. *Fallacies*. Advanced Reasoning Forum.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Willard Oliver. 1967. Formal fallacies and other invalid arguments. *Mind*, 76(304):463–478.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 644–657.

Frans H Van Eemeren and Rob Grootendorst. 1984. *Speech acts in argumentative discussions*. Dordrecht: Foris Publications.

Frans H Van Eemeren and Rob Grootendorst. 2016. *Argumentation, communication, and fallacies: A pragma-dialectical perspective*. Routledge.

Jacky Visser, John Lawrence, and Chris Reed. 2020. Reason-checking fake news. *Communications of the ACM*, 63(11):38–40.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Douglas N Walton. 1987. *Informal fallacies*, volume 4. John Benjamins Publishing.

Douglas N. Walton. 1995. *A pragmatic theory of fallacy*. Studies in rhetoric and communication. University of Alabama Press, Tuscaloosa.